Predviđanje visine godišnjih prihoda na temelju popisnih podataka

V. I. Banić

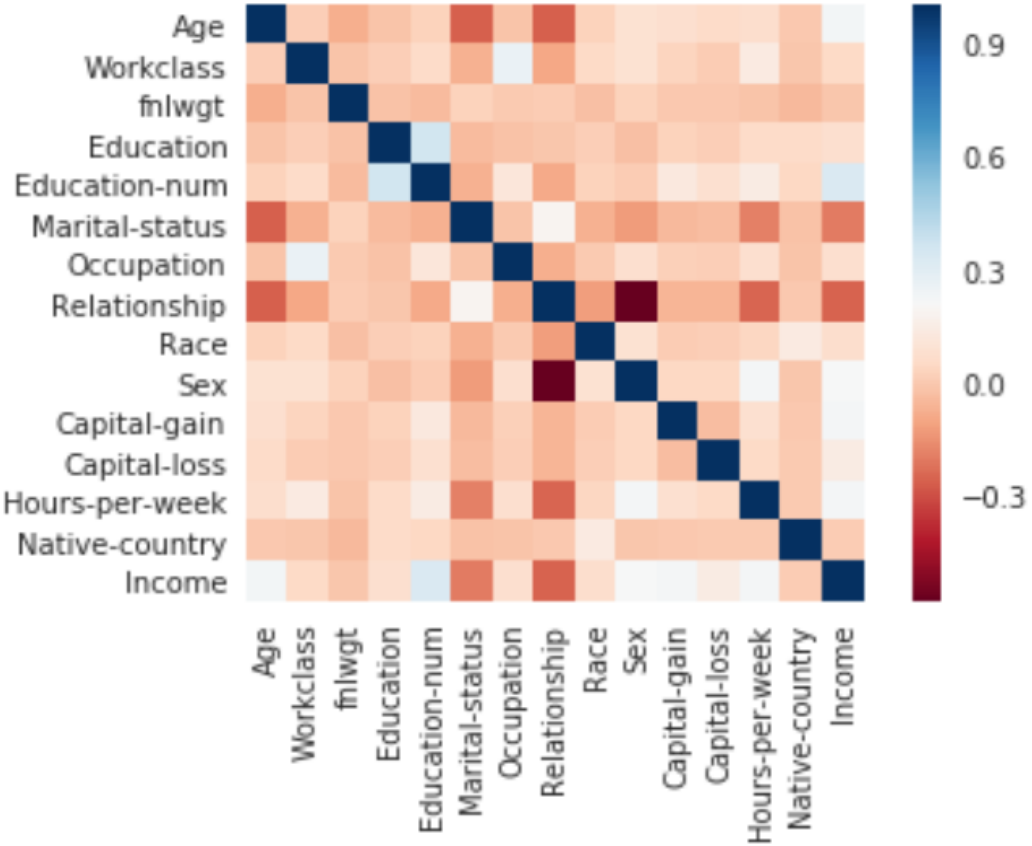
P. Bratulić

L. Bundalo

Opis problema

- Demografske i ekonomske karakteristike
- 14 atributa
- 48 842 instance

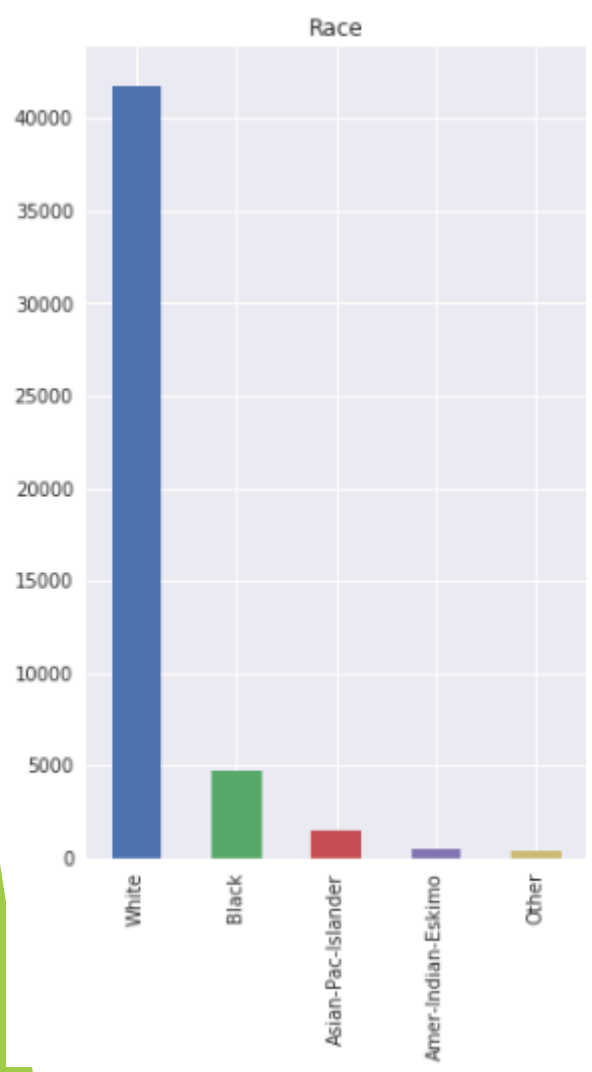
- **Age**: dob ispitanika, numerička varijabla
- **Workclass**: radni sektor, kategorijska varijabla (8 kategorija)
- **fnlwgt**: *final weight*, koristi se u anketama, ispitanici sa sličnim fnlwgt-om imaju slična demografska svojstva
- **Education**: najviši postignuti stupanj obrazovanja ispitanika, kategorijska varijabla (16 kategorija)
- **Education-num**: brojčana oznaka najvišeg postignutog stupnja obrazovanja ispitanika, numerička varijabla
- **Marital-status**: bračno stanje ispitanika, kategorijska varijabla (7 kategorija)
- **Occupation**: zanimanje ispitanika, kategorijska varijabla (14 kategorija)
- **Relationship**: položaj u obitelji ispitanika, kategorijska varijabla (6 kategorija)
- **Race**: rasa ispitanika, kategorijska varijabla (5 kategorija)
- **Sex**: spol ispitanika, kategorijska varijabla (2 kategorije)
- **Capital-gain**: prihod proizašao od investicija, numerička varijabla
- **Capital-loss**: gubitak proizašao od investicija, numerička varijabla
- **Hours-per-week**: broj radnih sati tjedno, numerička varijabla
- **Native-country**: država rođenja, kategorijska varijabla (41 kategorija)
- **Income**: podatak koji predviđamo (**Target**), zarađuje li ispitanik više ili manje od 50 000 američkih dolara tjedno, kategorijska varijabla (2 kategorije)

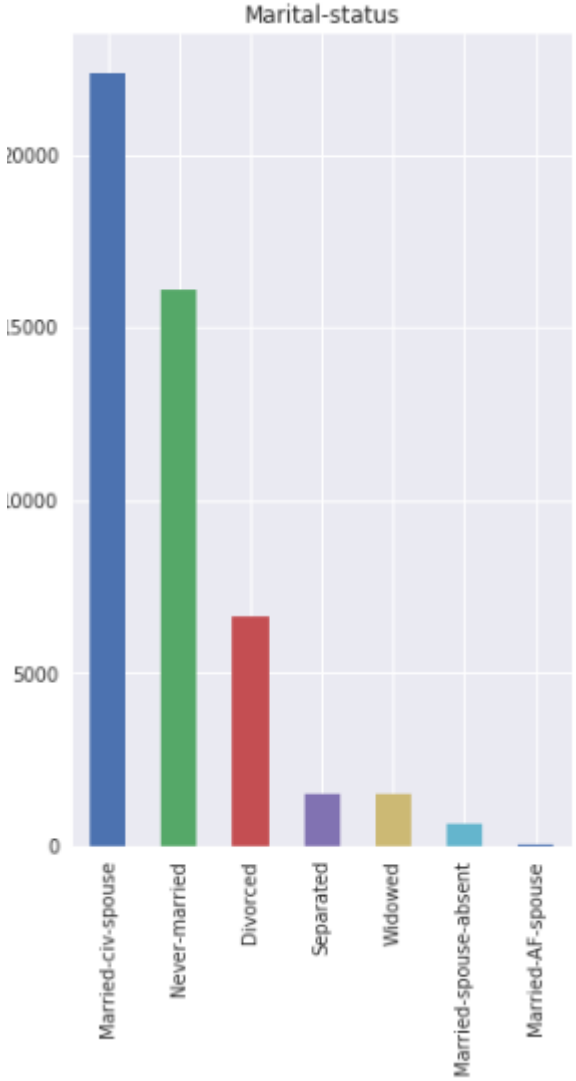

Age
Workclass
fnlwgt
Education
Education-num
Marital-status
Occupation
Relationship
Race
Sex
Capital-gain
Capital-loss
Hours-per-week
Native-country
Income
0.9
0.6
0.3
0.0
−0.3

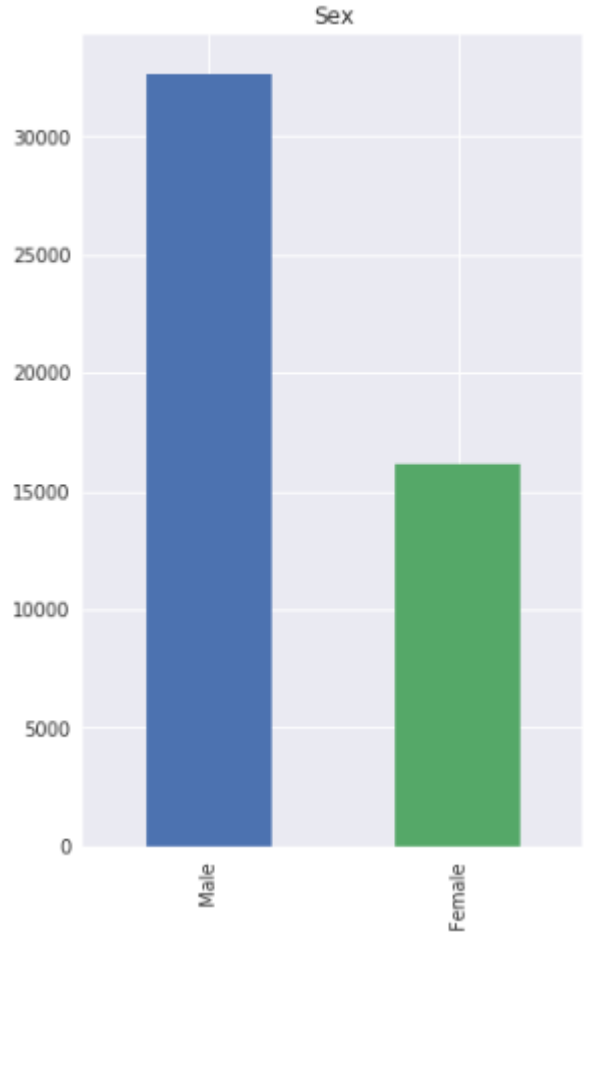
Distribucija podataka

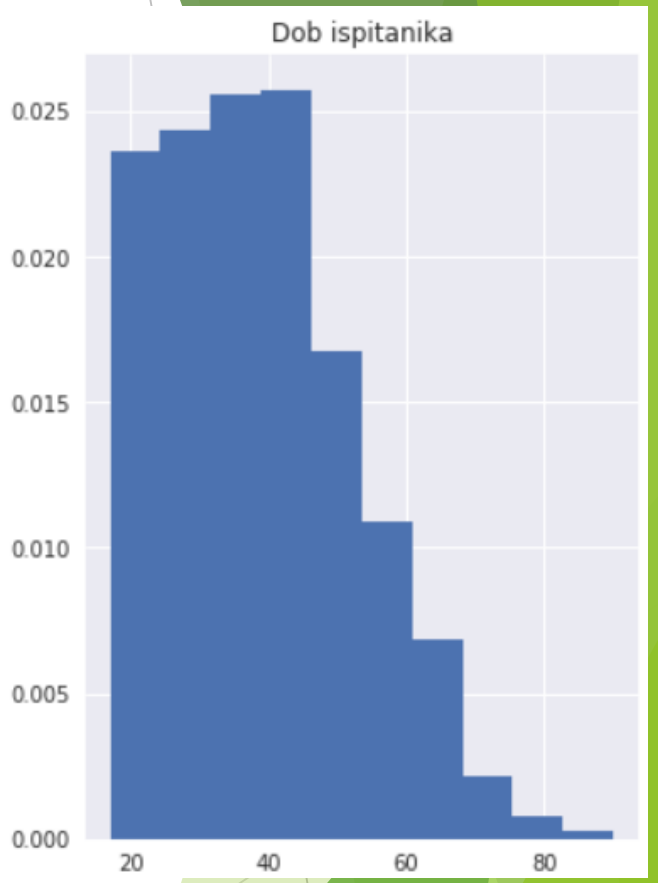
	Age	fnlwgt	Education-num	Capital-gain	Capital-loss	Hours-per-week
count	48842.000000	4.884200e+04	48842.000000	48842.000000	48842.000000	48842.000000
mean	38.643585	1.896641e+05	10.078089	1079.067626	87.502314	40.422382
std	13.710510	1.056040e+05	2.570973	7452.019058	403.004552	12.391444
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.175505e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.781445e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.376420e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.490400e+06	16.000000	99999.000000	4356.000000	99.000000

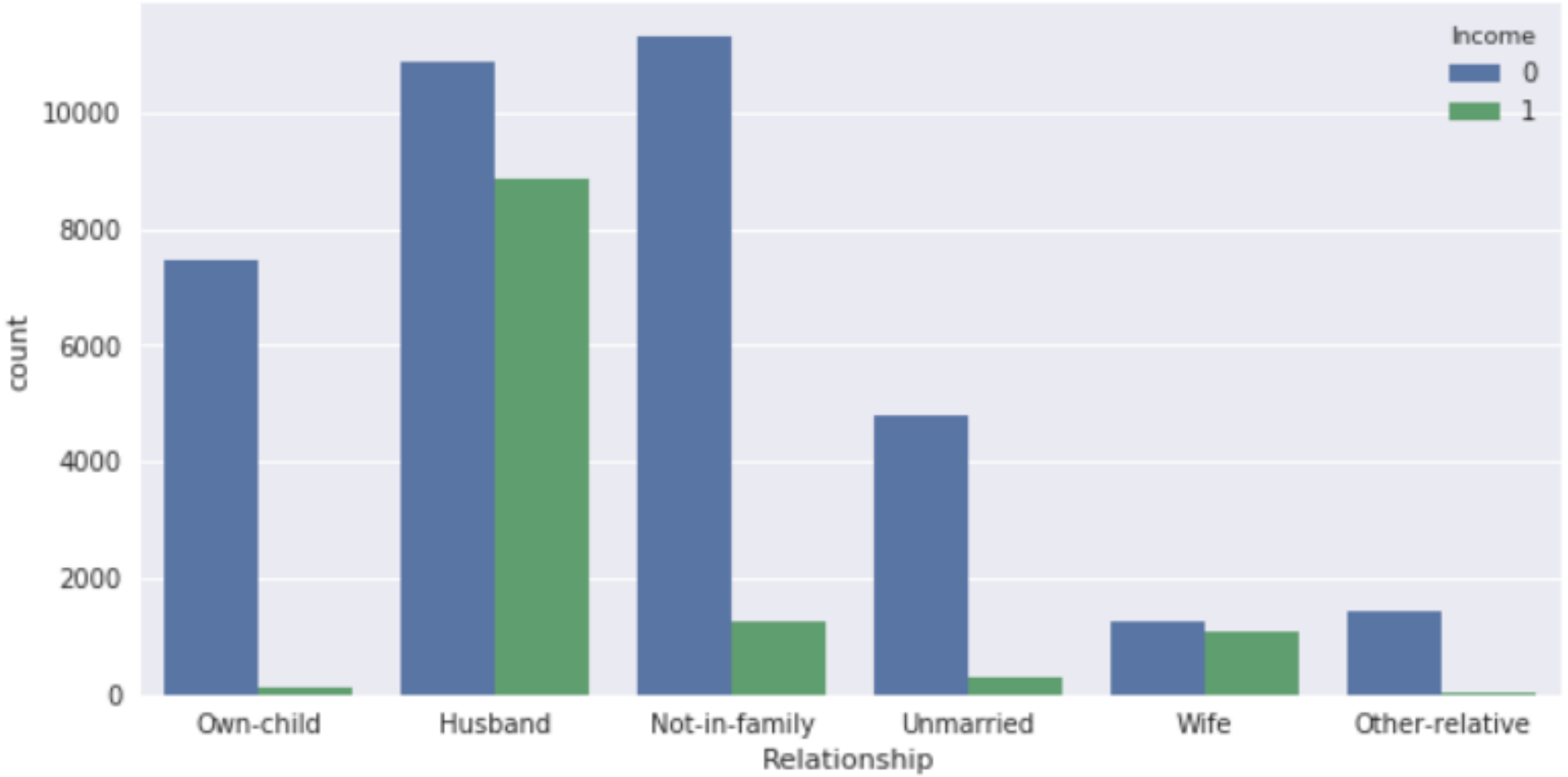
- Većina vrijednosti Capital-gain i Capital-loss iznosi 0.0

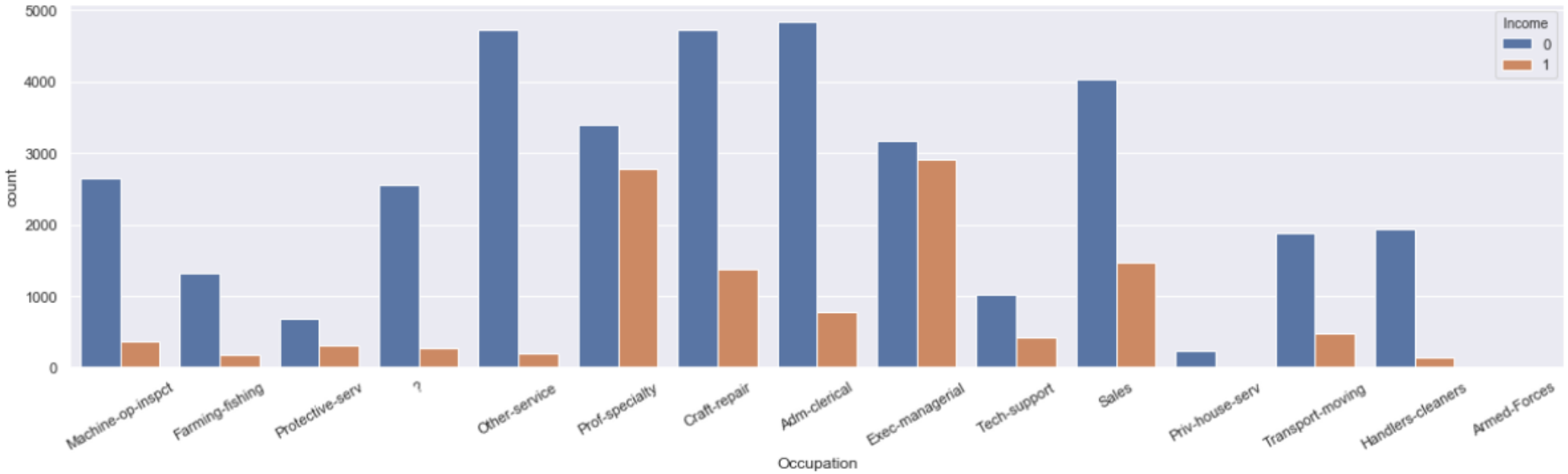

Race
40000
35000
30000
25000
20000
15000
10000
5000
0
White
Black
Asian-Pac-Islander
Amer-Indian-Eskimo
Other

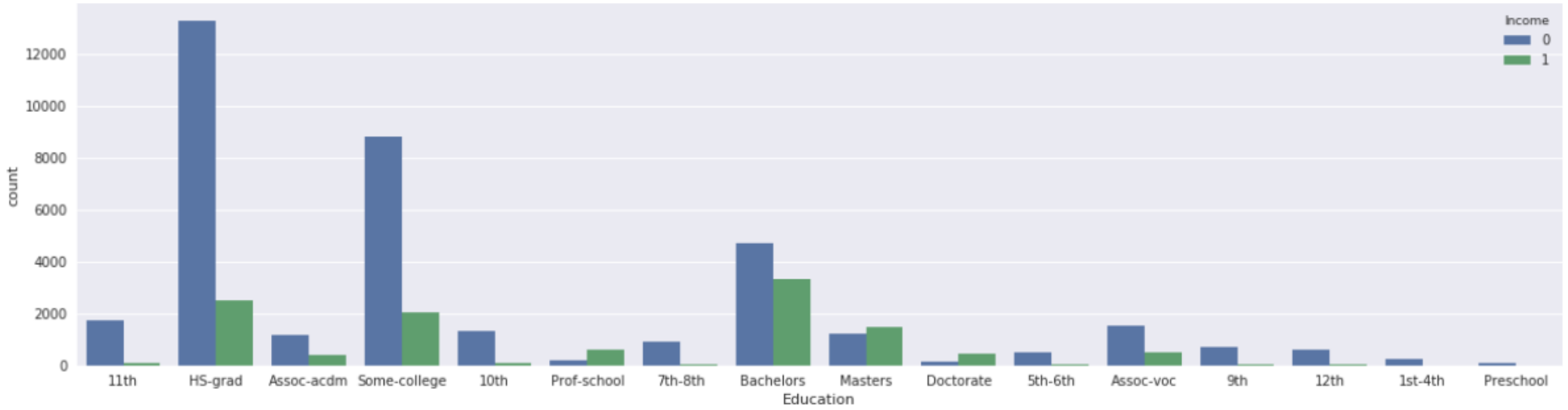

Marital-status
20000
15000
10000
5000
0
Married-civ-spouse
Never-married
Divorced
Separated
Widowed
Married-spouse-absent
Married-AF-spouse


Sex
30000
25000
20000
15000
10000
5000
0
Male
Female


Dob ispitanika
0.025
0.020
0.015
0.010
0.005
0.000
20
40
60
80


Income
0
1
10000
8000
6000
4000
2000
0
count
Own-child
Husband
Not-in-family
Unmarried
Wife
Other-relative
Relationship


5000
4000
3000
2000
1000
0
count
Income
0
1
Machine-op-inspct
Farming-fishing
Protective-serv
?
Other-service
Prof-specialty
Craft-repair
Adm-clerical
Exec-managerial
Tech-support
Sales
Priv-house-serv
Transport-moving
Handlers-cleaners
Armed-Forces
Occupation


Income
0
1
count
12000
10000
8000
6000
4000
2000
0
11th
HS-grad
Assoc-acdm
Some-college
10th
Prof-school
7th-8th
Bachelors
Masters
Doctorate
5th-6th
Assoc-voc
9th
12th
1st-4th
Preschool
Education

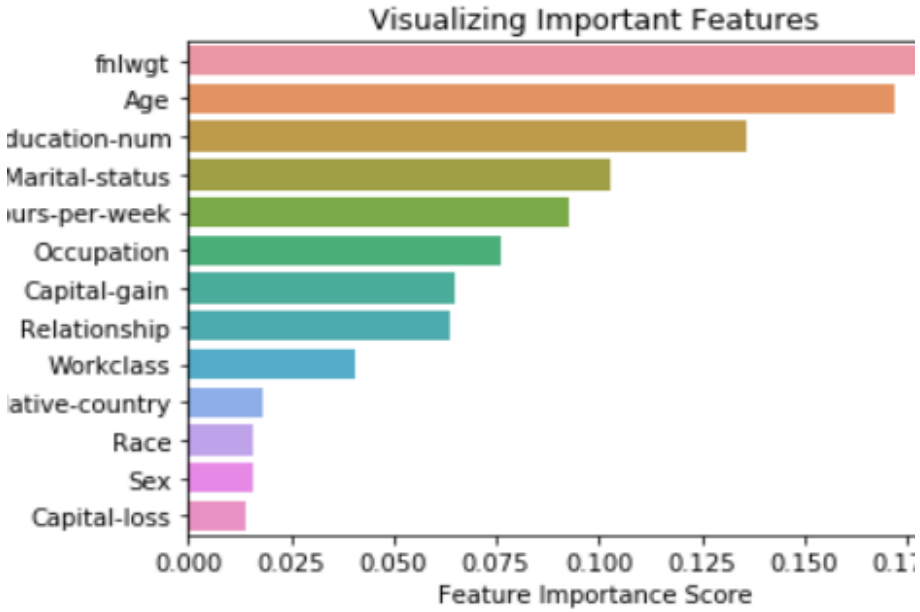
Opis rješavanja problema

- Algoritmi:
  - Logistička regresija
  - Random Forest
  - SVM
  - KNN algoritam
  - Naive Bayes
  - Neuronske mreže
- Mjera uspješnosti:
  - Točnost
  - Roc-auc-score

- Label encoding
- OneHot Encoding
- PCA
- StandardScaler
- RobustScaler
- MinMaxScaler

- Train-test-split

- Kategoričke vrijednosti → numeričke vrijednosti
- Nedostajuće vrijednosti → nova kategorija
- Izbacujemo Education
- Capital-gain i Capital-loss → kategorijske
- Smanjivanje dimenzionalnosti podataka



	Atribut	Occurance
0	Workclass	2799
1	Education	0
2	Marital-status	0
3	Occupation	2809
4	Relationship	0
5	Race	0
6	Sex	0
7	Native-country	857
8	Income	0

Logistička regresija

- Faktor regularizacije c
- Algoritam optimizacije
  - Liblinear
  - Newton-cg
- Točnost: 0.81
- Roc-auc-score: 0.83

Naive Bayes

- Točnost: 0.789
- Roc-auc-score: 0.806

Random forest

- Točnost: 0.844
- Roc-auc-score: 0.892

KNN

- Točnost: 0.734
- Roc-auc-score: 0.635

OneHot encoding

Logistička regresija

- Točnost: 0.855
- Roc-auc-score: 0.907
- Smanjena dimenzionalnost: 80 svojstava
  - Točnost: 0.846
  - Roc-auc-score: 0.899

Random forest

- Točnost: 0.839
- Roc-auc-score: 0.885
- Smanjena dimenzionalnost: 80 svojstava
  - Točnost: 0.84
  - Roc-auc-score: 0.886

Uklanjanje nedostajućih vrijednosti

- Logistička regresija
  - Bez poboljšanja
- Random forest
  - Lošiji rezultati

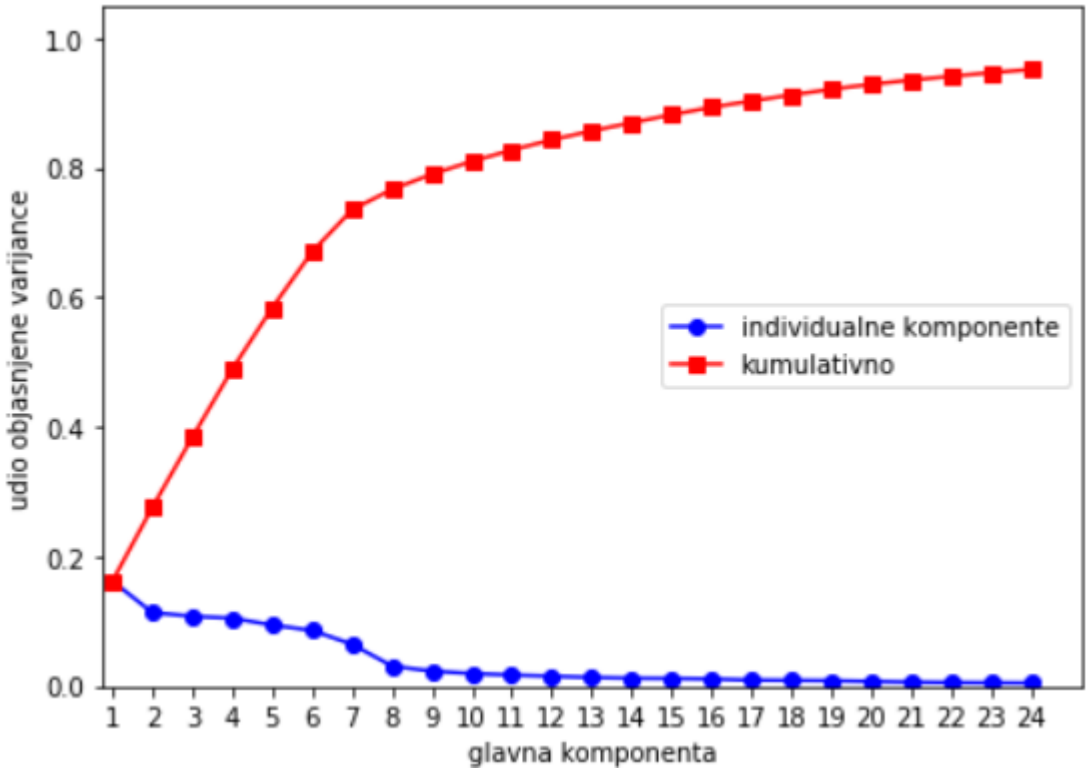
- Neuronske mreže
- 3 skrivena sloja, 13 neurona u sloju
- Točnost: 0.728
- Roc-auc-score: 0.887

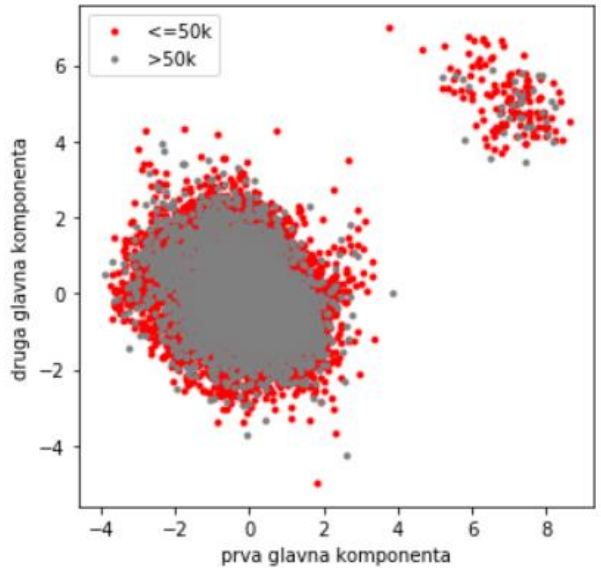
Micanje nedostajućih vrijednosti

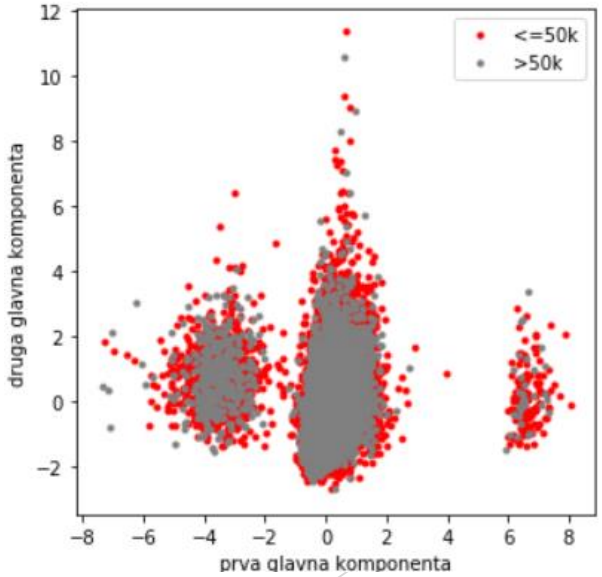
- Kategoričke vrijednosti → numeričke vrijednosti
- Izbacujemo Education
- Capital-gain i Capital-loss → kategorijske
- Skaliranje podataka (StandardScaler)
- Kategorijska svojstva koja nisu ordinalna → dummy-varijable (OneHot encoding)

PCA

- Odabrane glavne komponente daju 0.95 objašnjene varijance

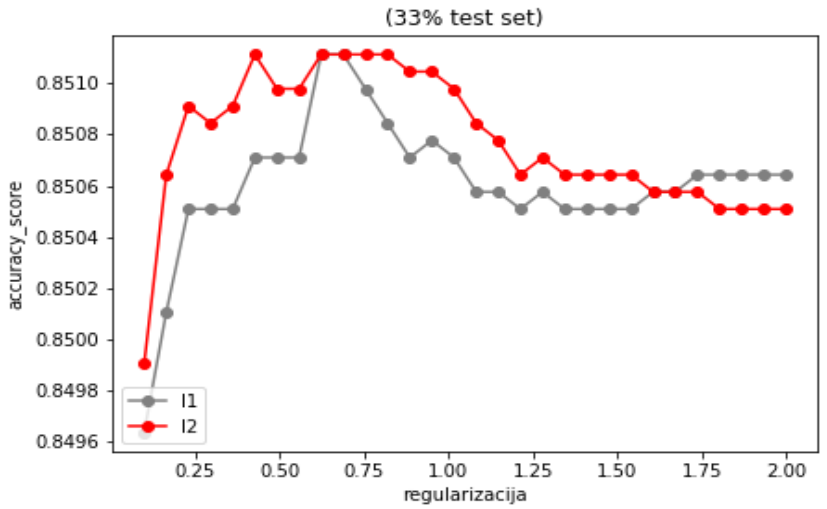


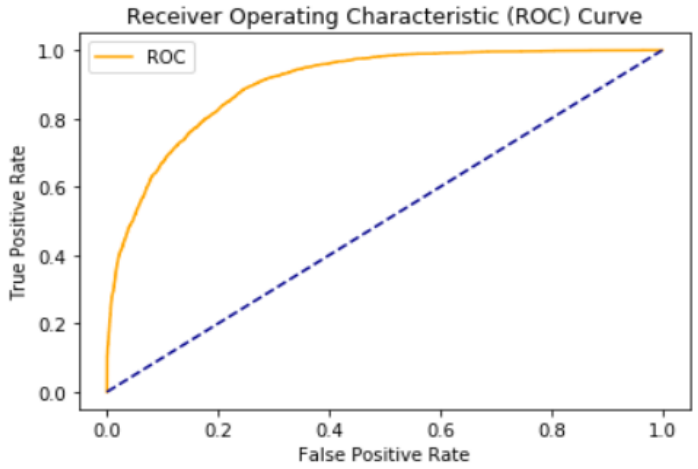


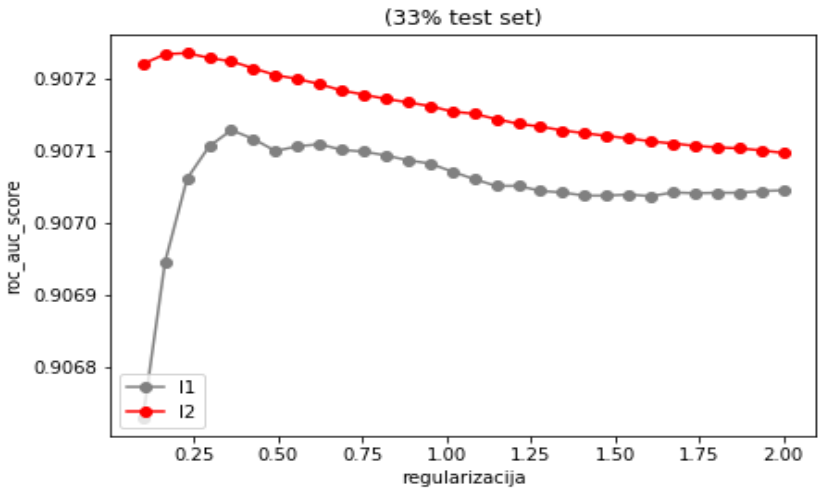


Logistička regresija

- Solver: newton-cg
- c = 0.8
- Točnost: 0.85
- Roc-auc-score: 0.91

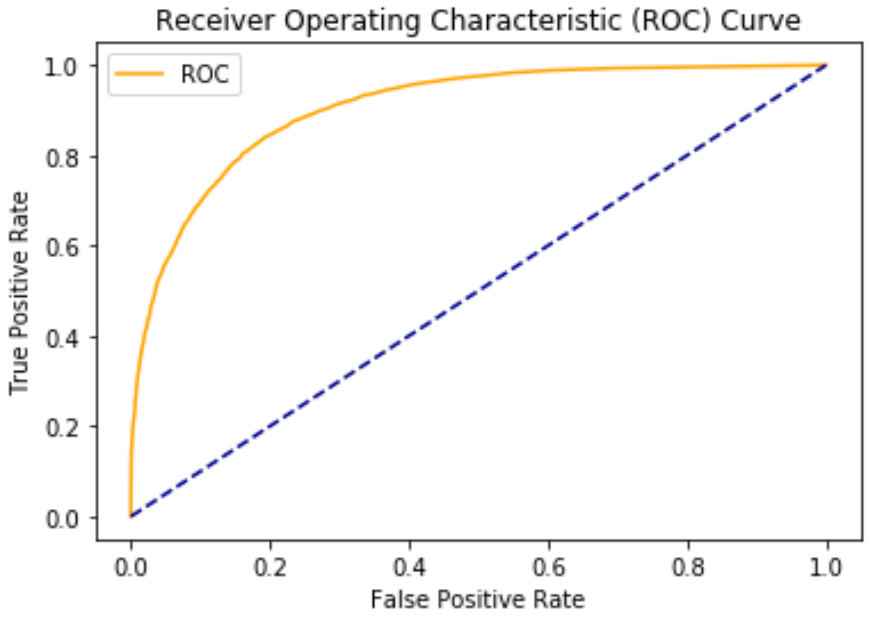






Random forest

- Bez PCA
  - Točnost: 0.856
  - Roc-auc-score: 0.906
- PCA:
  - Točnost: 0.842
  - Roc-auc-score: 0.896



SVM

- PCA
- linear kernel, c=0.3
  - Točnost: 0.85
  - roc-auc-score: 0.905
- kernel rbf, c=2.0
  - Točnost: 0.85
  - roc-auc score: 0.91

Naive Bayes

- PCA
  - Točnost: 0.792
  - roc-auc-score: 0.834

Neuronske mreže

- 3 sloja po 13 neurona
- Solver: sgd
- Točnost: 0.856
- roc-auc-score: 0.913

KNN

- Bez PCA
  - Točnost: 0.836
  - roc-auc-score: 0.867
- PCA
  - Točnost: 0.834
  - roc-auc-score: 0.864

RobustScaler

- Rezultati slični već dobivenim
- KNN bez PCA
  - Točnost: 0.854
  - roc-auc-score: 0.89
- Naive Bayes s PCA
  - Točnost: 0.791
  - roc-auc-score: 0.887

Promjene u skupu podataka

- Normalizacija numeričkih svojstva (StandardScaler)
- Kategorijska svojstva koja nisu ordinalna → dummy-varijable (OneHot encoding)
- Značajna korelacija svojstva ‘Relationship’ sa svojstvom ‘Marital-status’ → izbacujemo svojstvo ‘Relationship’
- 90% instanci ima vrijednost svojstva ‘Native-country’ USA → izbacujemo svojstvo ‘Native-country’
- Izbacujemo spol
- Diskretizacija normalnih vrijednosti

Usporedba rezultata

Naši najbolji rezultati

- Neuronske mreže (roc-auc-score 0.9127, točnost 0.8556)
- Logistička regresija (roc-auc-score 0.9071, točnost 0.8552)
- Random Forest (roc-auc-score 0.9079, točnost 0.8571)
- SVM (roc-auc-score 0.9047, točnost 0.8583)

Predicting earning potential on Adult Dataset

- KNN (roc-auc-score 0.889, točnost 0.8533)
- NBTree (roc-auc-score 0.908, točnost 0.8593)

Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid"

- NBTree (točnost oko 0.8590)
- Ostali algoritmi (točnost oko 0.84 ili 0.83)

Zaključak

- StandardScaler i SVM → najbolja točnost (0.8583)
- StandardScaler i neuronske mreže → najbolji roc-auc-score (0.9127)
- PCA → uglavnom bez poboljšanja
- Scaler → bolji rezultati

ROC_AUC

/	Log. Reg.	Log. Reg (PCA)	Random Forest	Random Forest (PCA)	SVM	SVM (PCA)	KNN	KNN (PCA)	Naive Bayes	Naive Bayes (PCA)	Neuronske mreže
No Scaling	0.9039	-	0.8909	-	-	-	0.6353	-	0.8055	-	0.8690
StandardScaler	0.9071	0.9026	0.8959	0.9054	0.9047	0.9005	0.8669	0.8642	0.8344	0.8586	0.9127
RobustScaler	0.9071	0.7921	0.8541	-	-	-	0.8542	0.8055	0.7906	0.7856	0.9121
MinMaxScaler	0.9071	0.8777	0.9079	0.8630	0.9043	0.8606	0.8505	0.8329	0.8568	0.8170	0.9102

Accuracy

/	Log. Reg.	Log. Reg (PCA)	Random Forest	Random Forest (PCA)	SVM	SVM (PCA)	KNN	KNN (PCA)	Naive Bayes	Naive Bayes (PCA)	Neuronske mreže
No Scaling	0.8552	-	0.8379	-	-	-	0.7337	-	0.7889	-	0.7946
StandardScaler	0.8508	0.8465	0.8571	0.8450	0.8583	0.8472	0.8303	0.8342	0.6235	0.7921	0.8556
RobustScaler	0.8509	0.7921	0.8565	0.8279	-	-	0.8542	0.8055	0.7906	0.7856	0.8544
MinMaxScaler	0.8504	0.8256	0.8542	0.8194	0.8486	0.8252	0.8261	0.8169	0.6062	0.74363	0.8525

Literatura

- [1] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/adult
- [2] http://robotics.stanford.edu/~ronnyk/nbtree.pdf
- [3] https://storage.googleapis.com/kaggle-forum-message-attachments/160002/5905/Paper%20on%20Machine%20Learning%20for%20Kaggle.pdf
- [4] http://robotics.stanford.edu/~ronnyk/nbtree.pdf Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996
- [5] https://towardsdatascience.com/normalization-vs-standardization-quantitative-analysis-a91e8a79cebf
- [6] https://scikit-learn.org/stable/index.html
- [7] http://www.dataminingmasters.com/uploads/studentProjects/Earning_potential_report.pdf
- [8] https://github.com/pmf-strojnoucenje/Vjezbe
- [9] https://scikit-learn.org/stable/modules/svm.html
- [10] https://scikit-learn.org/stable/modules/neighbors.html
- [11] https://scikit-learn.org/stable/modules/naive_bayes.html